 DOI: 10.1111/j.1541-0420.2008.01159.x

Differential Expression and Network Inferences through Functional Data Modeling

Donatello Telesca,[1] Lurdes Y.T. Inoue,[2,3] Mauricio Neira,[4] Ruth Etzioni,[3] Martin Gleave,[4,5] and Colleen Nelson[4,5]

[1]University of Texas, M.D. Anderson Cancer Center, Department of Biostatistics, Houston, Texas 77230, U.S.A.
[2]University of Washington, Department of Biostatistics, Seattle, Washington 98195, U.S.A.
[3]Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, U.S.A.
[4]The Prostate Centre at Vancouver General Hospital, Vancouver, British Columbia V6H 3Z6, Canada
[5]University of British Columbia, Department of Urologic Sciences, Vancouver, British Columbia V5Z 1M9, Canada

Summary. Time course microarray data consist of mRNA expression from a common set of genes collected at different time points. Such data are thought to reflect underlying biological processes developing over time. In this article, we propose a model that allows us to examine differential expression and gene network relationships using time course microarray data. We model each gene-expression profile as a random functional transformation of the scale, amplitude, and phase of a common curve. Inferences about the gene-specific amplitude parameters allow us to examine differential gene expression. Inferences about measures of functional similarity based on estimated time-transformation functions allow us to examine gene networks while accounting for features of the gene-expression profiles. We discuss applications to simulated data as well as to microarray data on prostate cancer progression.

Key words: Bayesian hierarchical model; Differential expression; Functional data; Functional similarity; Gene networks; Time course microarray data; Time transformation.

1. Introduction

Current research in molecular biology focuses on improving our understanding of gene regulation. Time course microarray data, consisting of mRNA expression from a common set of genes collected at different time points, provide new opportunities into the understanding of the gene regulation because it is believed that such data reflect underlying biological processes developing over time.

Graphical models and, in particular, Bayesian networks have been largely utilized to study gene regulation using cross-sectional microarray data (see, e.g., Markowetz and Spang [2007] and the references therein). Dynamic Bayesian networks have been applied to time course microarray data as they extend Bayesian networks by allowing cyclic temporal relationships between genes. Although appealing, dynamic Bayesian networks have computational limitations because complexity grows quickly with the number of genes. Moreover, time delays and/or dynamic changes of the network have mostly been addressed within a simplified view to reduce the computational burden. Some authors, for example, analyzed gene networks assuming that relationships were linear and time homogeneous (see, e.g., Beal et al., 2005; Inoue et al., 2007). Opgen-Rhein and Strimmer (2006a) proposed an extension of the graphical models to the dynamic setting by treating the observed time course expression data as functional data and proposing a partial correlation measure of dependence between any pair of coexpressed gene-expression profiles.

There is a large body of evidence supporting the idea that coexpressed genes are more likely to be coregulated (Allocco, Kohane, and Butte, 2004; Michalak, 2008). This idea has been expanded to allow for time delays. Time-delayed expression profiles are associated with a series of biological events such as the cell cycle, circadian clock, cell differentiation, and development (Weber, Kramer, and Fussenegger, 2007). In fact, Bratsun et al. (2005) observe that the modeling of time delays provides an approximation to modeling a complex sequence of biochemical events underlying transcription and translation of any gene.

Some authors have explored the temporal structure of the expression profiles. Qian et al. (2001) use dynamic programming to obtain alignment of the expression profiles of any pair of genes and identify time-delayed activation or inhibitory relationships. This approach is, however, based on alignment scores obtained from the raw data, which may be problematic with microarray data because the signal-to-noise ratio is often very small. In the context of time ordering, Leng and Müller (2006) use a model-based approach, estimating the time shift for gene profiles to obtain an optimal pairwise alignment. While this procedure accounts for variability in the observed mRNA intensity, the assumption of a strictly linear time shift may be inappropriate when the mRNA abundance signal exhibits multiple features in its profile over time.

We propose a model that allows us to investigate the dynamics of gene relationships. Our method relies on the extraction of information about the timing of features, such as peaks

and valleys, in each gene-expression profile. Specifically, gene-expression profiles are modeled as realizations of a compound process involving a random transformation of a common profile and a transformation of the timing of the features of the profile. Unlike previous approaches, our model allows for a broader class of relationships with possible nonlinear time transformations and does not require equally spaced sampling or presmoothed trajectories. The model builds on Telesca and Inoue (2008) who extended the classical self-modeling regression models (Ramsay and Li, 1998; Brumback and Lindstrom, 2004; Gervini and Gasser, 2004) by using a Bayesian hierarchical modeling approach. In this article, we discuss model-based selection of differentially expressed genes and describe a probabilistic framework for the investigation of regulatory relationships between genes. We propose measures of association, in particular, assessing dynamic network relationships using timing maps. We show through a case study that our method validates many relationships currently supported by the literature.

The remainder of this article is organized as follows. In Section 2, we describe our model and inferences about differential expression and gene network. In Section 3, we apply our model to simulated data and to a time course gene-expression microarray dataset from animal experiments on the progression of prostate cancer. Finally, in Section 4, we provide a discussion.

2. Model Formulation

2.1 *Model Description*

Let $y_i(t)$ denote the observed expression level of gene i at time t where $i = 1, 2, \ldots, N$ and $t \in T = [t_1, t_n]$. We introduce the following three-stage hierarchical model.

Stage One: The observed value of the trajectory of gene i at time t is:

$$y_i(t) = c_i + a_i m\{u_i(t, \boldsymbol{\phi}_i), \boldsymbol{\beta}\} + \epsilon_i, \quad i = 1, \ldots, N, \quad t \in T, \tag{1}$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_\epsilon)$.

In the above, $u_i(\cdot, \cdot)$ denotes the gene-specific time-transformation function and $m(\cdot, \cdot)$ denotes a common shape function generating the individual trajectories. We use flexible representations of both functions using B-splines (de Boor, 1978). Specifically, the curve-specific random time-transformation functions characterizing the timing features of each curve are defined as $u_i(t, \boldsymbol{\phi}_i) = \mathcal{B}'_u(t)\boldsymbol{\phi}_i$, where $\mathcal{B}_u(t)$ is a set of B-spline basis and $\boldsymbol{\phi}_i$ is a $Q-$dimensional vector of basis coefficients. We define u_i as a smooth monotone map over the design interval T with values on a compact interval $\mathcal{T} = [t_1 - \Delta, t_n + \Delta]$ where $\Delta \geq 0$. To ensure monotonicity and a boundary on the image of these functions, we impose constraints on the time-transformation coefficients $\boldsymbol{\phi}_i$, namely,

$$(t_1 - \Delta) < \phi_{i1} < \cdots < \phi_{iq} < \phi_{i(q+1)} < \cdots < \phi_{iQ} < (t_n + \Delta), \tag{2}$$

$$\phi_{i1} \in [(t_1 - \Delta), (t_1 + \Delta)], \quad \phi_{iQ} = t_n + \phi_{i1}, \tag{3}$$

for all genes $i = 1, \ldots, N$.

Similarly, we represent $m\{u_i(t, \boldsymbol{\phi}_i), \boldsymbol{\beta}\} = \mathcal{B}'_m\{u_i(t, \boldsymbol{\phi}_i)\}\boldsymbol{\beta}$, where $\mathcal{B}_m\{u_i(t, \boldsymbol{\phi}_i)\}$ is a set of B-spline basis functions and $\boldsymbol{\beta}$ is a $K-$dimensional vector of basis coefficients. To ensure that $\mathcal{B}_m\{u_i(t, \boldsymbol{\phi}_i)\}$ spans a functional space over the extended design interval $\mathcal{T}$, the common shape function is defined so that $m(\cdot, \cdot) : \mathcal{T} \longrightarrow \mathbb{R}$.

Stage Two: Given a common shape function $m(\cdot, \cdot)$, individual curves may exhibit different levels and amplitudes of response. We assume that the gene-specific level $c_i \overset{iid}{\sim} \mathcal{N}(c_0, \sigma^2_c)$. Parameter a_i describes the amplitude of the mRNA signal for gene i. We formalize our statistical definition of differentially expressed genes via a mixture approach. Our approach is similar to that presented by Parmigiani et al. (2002). For each gene, we specify the following prior for the amplitude of the expression signal,

$$\begin{aligned} a_i = {} & \pi^- N\left(a_0^-, \sigma^2_{a^-}\right) I(a_i < 0) + \pi^+ N\left(a_0^+, \sigma^2_{a^+}\right) I(a_i > 0) \\ & + \pi^0 N\left(0, \sigma^2_{a^0}\right), \quad i = 1, \ldots, N, \end{aligned} \tag{4}$$

with $(\pi^- + \pi^0 + \pi^+) = 1$. Here π^0 identifies the overall proportion of genes in their normal range of variation, while $(\pi^- + \pi^+)$ identifies the proportion of overly active genes. The mixture characterization with two truncated normals (i.e., $N^-(\cdot, \cdot)\, I(a_i < 0)$ and $N^+(\cdot, \cdot)\, I(a_i > 0)$) allows us to account for genes with a synchronous expression signal of opposite sign (negative dependence).

We model the time-transformation function coefficients as following a multivariate normal distribution $\boldsymbol{\phi}_i \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\Upsilon}, \boldsymbol{\Sigma}_\phi)$, where $\boldsymbol{\Upsilon}$ is the vector associated with the identity time-transformation function so that $u_i(t, \boldsymbol{\Upsilon}) = t$.

Stage Three: We assume that $a_0^+ \sim \mathcal{N}(1, \sigma^2_{a0}), a_0^- \sim \mathcal{N}(-1, \sigma^2_{a0})$, and $c_0 \sim \mathcal{N}(0, \sigma^2_{c0})$. Moreover, $1/\sigma^2_{a+}, 1/\sigma^2_{a-}, 1/\sigma^2_{a^0} \sim \mathcal{G}(a_a, b_a)$. In particular, to accommodate heavy tails in the genomic distribution of mRNA abundance we require $\sigma^2_{a^0} < \min(\sigma^2_{a^-}, \sigma^2_{a^+})$. Finally, we assume that $1/\sigma^2_c \sim \mathcal{G}(a_c, b_c)$, and $1/\sigma^2_\epsilon \sim \mathcal{G}(a_\epsilon, b_\epsilon)$. (In our formulation, $X \sim \mathcal{G}(a, b)$ indicates a *Gamma* distribution, parameterized so that $E(X) = a/b$). The mixture proportions $\boldsymbol{\pi} = (\pi^+, \pi^-, \pi^0)'$ have a conjugate Dirichlet prior $\mathcal{D}(\boldsymbol{\alpha})$.

Additionally, we assume that the shape function coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ follow a second-order shrinkage process (Eilers and Marx, 1996). Thus, we model $\beta_\kappa = 2\beta_{\kappa-1} - \beta_{\kappa-2} + \eta_\kappa$, with $\eta_\kappa \sim \mathcal{N}(0, \lambda)$ and $1/\lambda \sim \mathcal{G}(a_\lambda, b_\lambda)$. Similarly, for the time-transformation parameters we use a first-order shrinkage process so that $(\phi_{iq} - \Upsilon_q) = (\phi_{iq-1} - \Upsilon_{q-1}) + \nu_{iq}$, with $\nu_{iq} \sim \mathcal{N}(0, \sigma^2_\phi)$ and $1/\sigma^2_\phi \sim \mathcal{G}(a_\phi, b_\phi)$.

2.1.1 *Choosing priors.* For practical implementation of the model, using normalized mRNA data, we assume that the prior distribution of c_i is concentrated between min $(\mathbf{Y})$ and max $(\mathbf{Y})$. Similarly, the absolute amplitude of expression $|a_i|$, is centered around 1 and may range between 0 and 10. Given the above domains of c_i and a_i, then assuming a $\mathcal{G}(0.1, 1)$ for the precision parameters $1/\sigma^2_a$ and $1/\sigma^2_c$ implies relatively diffuse priors. When choosing a prior for the time-transformation coefficients, we note that the natural domain of the parameters ϕ_i is constrained to the interval $(t_1 - \Delta,\ t_n + \Delta)$. Rescaling the above interval to the (0, 1) interval, we assume that $1/\sigma^2_\phi \sim \mathcal{G}(0.01, 100)$ which is also relatively diffuse on the rescaled interval. Finally, the choice of Δ depends on the

application. In our simulation study, we used $\Delta < 5$ with the upper bound reflecting approximately the periodicity in the simulated curves. In the case study we used $\Delta = 7$, which biologically corresponds to the time period when the tumor starts to regrow.

Sensitivity analysis to our prior choices is presented in the Web Supplementary Materials, Section 1. Our analysis indicate that the above priors are fairly noninformative.

2.1.2 *Choosing spline basis, location, and number of knots.* Our model depends on specific choices for the spline basis, the location and the number of spline knots modeling the common shape function $m(t, \boldsymbol{\beta})$, and the individual time-transformation functions $\mu_i(t, \boldsymbol{\phi}_i)$.

We consider B-spline basis of order 4, because of their numerical stability (Peña, 1997). Also they allow for a simple translation of functional constraints (monotonicity and image) into constraints over the basis coefficients as represented by equations (2) and (3).

There are some practical considerations regarding the number of spline knots used to model the shape and the time-transformation functions. When modeling the common shape function, we borrow information from the entire set of profiles. In our applications, using the number of knots equal to the number of sampling time points provides great modeling flexibility. Moreover, the shrinkage process on the basis coefficients (as described in Section 2.1) allows for adaptive smoothing and makes our inferences less dependent on the chosen number of knots (see Supplementary Materials, Section 3). Different considerations apply when we model the individual time-transformation functions. These functions carry structural smoothness as they are constrained to be monotone. This requirement counterbalances the small number of observations associated with each gene profile and suggests parsimony in the choice of the number of knots. In our applications a number of knots between 3 and 6 allowed for enough flexibility (see Web Supplementary Materials, Section 2). Finally, because in our formulation the time scale is stochastic, the knots are equally spaced.

2.2 *Estimation and Inference*

Let $\boldsymbol{\theta}$ denote the full parameter vector, that is, $\boldsymbol{\theta} = (\mathbf{c}', \mathbf{a}', \boldsymbol{\beta}', \boldsymbol{\phi}', \boldsymbol{\pi}', c_0, a_0, \sigma^2_\epsilon, \sigma^2_c, \sigma^2_{a0}, \sigma^2_{a-}, \sigma^2_{a+}, \lambda, \sigma^2_\phi)'$, where $\mathbf{c} = (c_1, \ldots, c_N)'$, $\mathbf{a} = (a_1, \ldots, a_N)'$ and $\boldsymbol{\phi} = (\boldsymbol{\phi}'_1, \ldots, \boldsymbol{\phi}'_N)'$ is an $N \times Q$ vector of individual time-transformation parameters. We fully specify the Bayesian model with priors on the parameter vector $\boldsymbol{\theta}$ as discussed in Section 2.1.

The joint posterior density of $\boldsymbol{\theta}$ conditional on data $\mathbf{Y}$ is analytically intractable, and so we implemented a Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution. Specifically, we use Metropolis–Hastings to sample the time-transformation parameters $\boldsymbol{\phi}$ and Gibbs sampling steps to sample the remaining parameters for which the full conditionals are available in closed form. Updating of the amplitude parameters $\mathbf{a}$ is based on augmented data with the set of mixture class indicators $\mathbf{z} = (z_1, \ldots, z_N)'$ for all genes.

Our inferences are based on examining and postprocessing the MCMC samples from the posterior distribution of $\boldsymbol{\theta}$. Next, we discuss inferential analysis from our model. The goal is to make inferences about interactions among a set of differentially expressed genes. We can address this problem in two steps. First, we select differentially expressed genes, which in our applications we define as genes that do not have a constant level of mRNA over time. Second, we proceed with the analysis of interactions between differentially expressed genes using timing maps.

2.2.1 *Differential expression.* Assessment of differential expression using time course data has been studied under the frequentist or the Bayesian paradigm. Specifically, expression profiles are usually modeled using linear combinations of orthonormal basis (Angelini et al., 2007; Storey, 2007) and differential expression is defined as a significant variation of the mRNA abundance signal over time. The issue of multiple testing is addressed considering adjustments for familywise error rates, either via resampling techniques (Storey, 2007) or via Bayesian hierarchical adjustments (Angelini et al., 2007; Chi et al., 2007).

In the context of the model described in Section 2.1, we start by observing that the amplitude parameter vector $\boldsymbol{a} = (a_1, \ldots, a_N)'$ is informative about the strength of the mRNA signal. Thus, we can use it to identify differentially expressed genes. Specifically, we address this question using the following set of hypotheses:

$$\begin{aligned} &H_{0i} : a_i \sim N\left(0, \sigma^2_{a^0}\right) \text{ versus} \\ &H_{1i} : a_i \sim N\left(a^+_0, \sigma^2_{a^+}\right) \text{ or } a_i \sim N\left(a^-_0, \sigma^2_{a^-}\right); \\ &\quad i = 1, \ldots, N. \end{aligned} \tag{5}$$

When testing a large number of hypotheses it is desirable to control for some predefined error rate. A popular choice is to control the false discovery rate (FDR; Benjamini and Hochberg, 1995). Following Müller, Parmigiani, and Rice (2006), for a given null hypothesis H_{0i}, let $\delta_i = I(\text{Reject } H_{0i})$ be the indicator for the decision about H_{0i}, $D = \sum_i \delta_i$ denote the total number of rejections, and $r_i = I(H_{0i} \text{ False})$ denote the indicator of the unknown truth. The FDR is $FDR = \sum_i \delta_i(1 - r_i)/D$. Under the Bayesian approach, because r_i is unknown, we could control the expected posterior FDR. Defining $v_i = P(r_i = 1 \,|\, \mathbf{Y})$, the expected posterior FDR is given by:

$$E(FDR \mid \mathbf{Y}) = \sum_i (1 - v_i)\delta_i / D. \tag{6}$$

Newton et al. (2004) and Morris et al. (2006) apply this idea considering rules that reject H_{0i} if $v_i > \gamma^*$, where γ^* is selected so that the expected posterior FDR is controlled at a given level α.

The choice of a decision rule can be formalized with the specification of loss functions. In fact, Müller et al. (2004) provide several examples of loss functions that induce decision rules of the form $\delta_i = I(v_i > \gamma^*)$. A disadvantage of the loss functions inducing the above decision rule is that they do not fully account for the expression levels. Müller et al. (2006) propose an alternative loss function, which in the context of our model is:

$$\mathcal{L}(a, \delta) = K \sum_i (1 - \delta_i)|a_i| - \sum_i \delta_i |a_i| + \varsigma D, \tag{7}$$

where K is the tradeoff between rejecting or not the null hypothesis and ς is the cost associated with rejecting H_0. The above loss function implies that the optimal decision rule is:

$$\delta_i^* = I\left\{E(|a_i| \mid \mathbf{Y}) = \bar{\mathbf{m}}_i > \frac{\varsigma}{1+K}\right\}. \quad (8)$$

In our applications, we consider a combined criteria accounting for the strength of evidence in the amplitude while controlling for the expected posterior FDR. Specifically, we consider decision rules provided by equation (8), choosing ς such that $E(FDR\,|\,\mathbf{Y}) \leq \alpha$. Defining $p_i = (1 - v_i)$, it can be easily shown that the optimal cost ς^* that explicitly controls for the FDR is $\varsigma^* = (1+K)\,\bar{\mathbf{m}}_\ell$, with $\ell = \sup(i : \sum_{j=1}^{i} p_j \leq i\alpha)$ and p_j ordered so that $\bar{\mathbf{m}}_1 \geq \bar{\mathbf{m}}_2 \geq \ldots \geq \bar{\mathbf{m}}_N$.

2.2.2 *Network inferences.* The underlying idea for the investigation of gene networks using time course microarray data is that genes that share similar expression profiles may share similar biological functions and thus, could be related. Three aspects are, however, not always collectively taken into account by traditional network models. First, that genes often exhibit different levels and different changes in amplitude of their mRNA abundance despite being related. Second, that relationships may be time delayed as seen, for example, between transcription factors and their targets. And, third, that relationships may have a dynamic aspect changing over time.

This motivates our work. We investigate relationships between genes accounting for gene-specific patterns of expression. We assume that two genes are related if their expression profiles, up to scale, have similar timing features. To illustrate this idea, consider the profiles for two hypothetical genes in panel (a) of Figure 1. Features such as peaks and valleys in the profile shown in solid line (gene A) are delayed in relation to those observed in dashed line (gene B). The corresponding time-transformation functions in panel (b) highlight the time shift. Because for all time points the time-transformation functions show that timing features of gene B anticipate that of gene A, they are suggestive that gene B has a regulatory effect over gene A. Panel (c) shows another example where looking at the profiles alone may indicate that there is no relationship between two genes. Here, the two profiles have an overall small correlation (correlation = −0.31), indicating

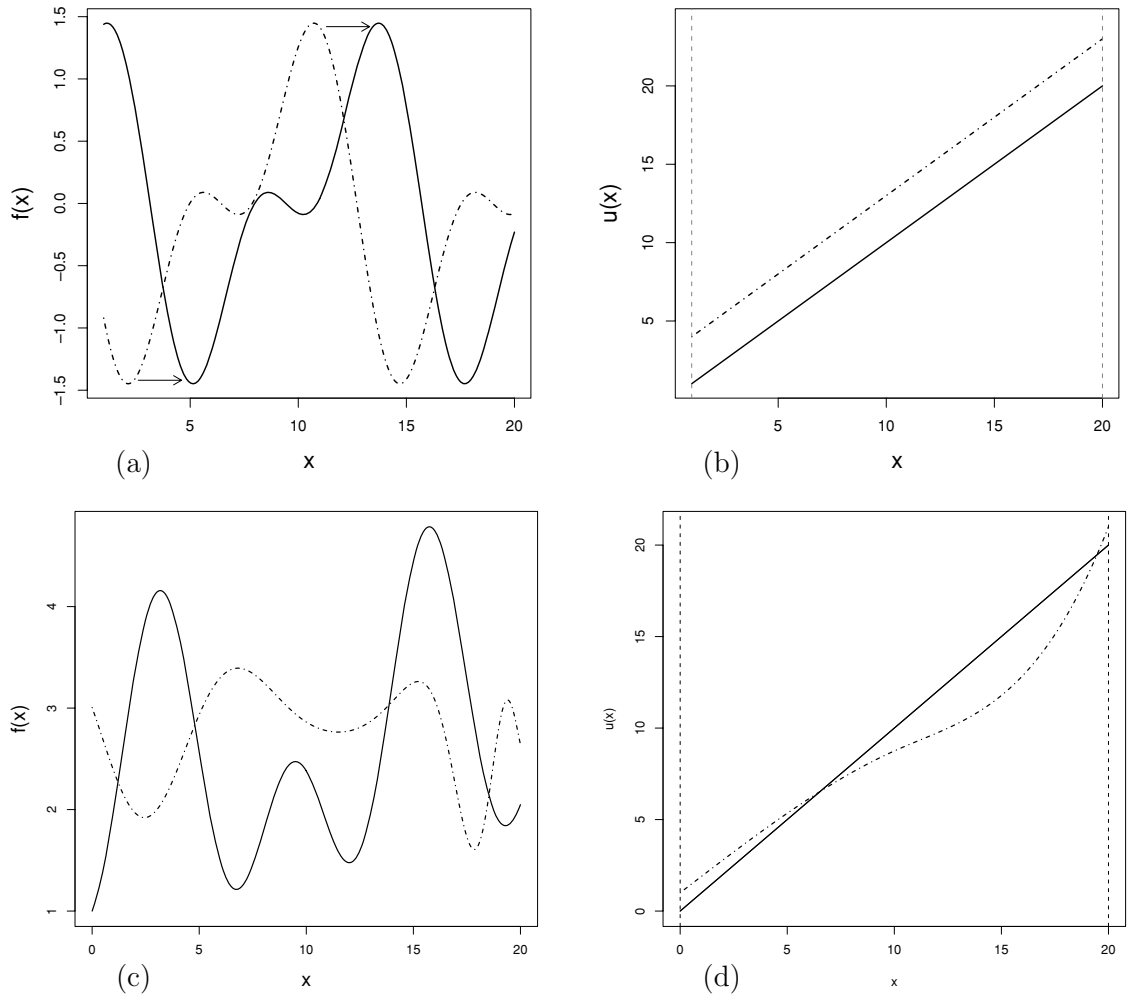

Figure 1. Motivating example. Panels (a) and (c): Profile for two hypothetical genes (gene A in full line, gene B in dashed line). Profiles are derived from composite functions $f_i(x) = m(u_i(x))$. Panels (b) and (d): Time-transformation functions $u_i(x)$ describing the timing of profile features (from profiles shown in panels (a) and (c), respectively).

no relationship. However, the time-transformation function in panel (d) is very informative about the dynamic similarity of the two profiles. In particular, we notice that the two profiles are fairly synchronized in the first half of the design interval, but much less so in the second half.

We thus propose using the time-transformation functions to derive measures of relationships that are based on functional similarities.

DEFINITION: *We define a local distance* $d_{ik}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_k, t)$ *between genes* i *and* k $(i \neq k)$ *with* $t \in [t_1, t_n]$ *as*

$$d_{ik} = d_{ik}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_k, t) = |u_i(t, \boldsymbol{\phi}_i) - u_k(t, \boldsymbol{\phi}_k)|, \tag{9}$$

that is, as the absolute distance between the time-transformation functions of genes i *and* k *at time* t*. The local distance may be interpreted as the time shift between the expression profile features of two genes at a given time point.*

One may adapt the above local distance by looking at the network in subsets of the sampling design. In the more extreme end where we look at the network over the entire observation period we can define a global distance measure as follows.

DEFINITION: *We define a global distance* $D_{ik}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_k)$ *summarizing the pairwise profile similarity between genes* i *and* k *as*

$$D_{ik} = D_{ik}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_k) = \sum_{j=1}^{n} |u_i(t_j, \boldsymbol{\phi}_i) - u_k(t_j, \boldsymbol{\phi}_k)| \,/(t_n - t_1), \tag{10}$$

that is, as the average absolute distance between the time-transformation functions evaluated on the time points of the sampling design. The global distance can be interpreted as the average distance between the timing of the curve features characterizing the expression profiles of two genes.

Recall that our inferences are based on samples from the posterior distribution of the model parameters. Let $\boldsymbol{\phi}_i^{(j)}$ denote the jth draw from the marginal posterior distribution of the time-transformation coefficient $\boldsymbol{\phi}_i, i = 1, \ldots, N;\ j = 1, \ldots, M$. Draws from the marginal posterior distribution of the time-transformation function $u_i(t, \boldsymbol{\phi}_i) = \mathcal{B}'_u(t)\boldsymbol{\phi}_i$ at time t are given by:

$$u_i^{(j)}(t, \boldsymbol{\phi}_i) = \mathcal{B}'_u(t)\boldsymbol{\phi}_i^{(j)}, \quad j = 1, \ldots, M. \tag{11}$$

For all pairs of genes $i \neq k$, we can then derive the marginal posterior distributions of the pairwise local and global distances by applying equations (9) and (10) to the samples in equation (11) so that:

$$\begin{aligned} d_{ik}^{(j)} &= \left|u_i^{(j)}(t, \boldsymbol{\phi}_i) - u_k^{(j)}(t, \boldsymbol{\phi}_k)\right|, && j = 1, \ldots, M; \\ D_{ik}^{(j)} &= \sum_{j=1}^{n} \left|u_i^{(j)}(t_j, \boldsymbol{\phi}_i) - u_k^{(j)}(t_j, \boldsymbol{\phi}_k)\right| \big/ (t_n - t_1), && j = 1, \ldots, M. \end{aligned} \tag{12}$$

Relevant summaries from the marginal distributions may be extracted to draw conclusions on the relationships. In particular, given the expected posterior distances $E(D_{ik} \mid \mathbf{Y}) \approx 1/M \sum_{j=1}^{M} D_{ik}^{(j)}$, we can use a decision-theoretic formulation and select gene pairs satisfying $E(D_{ik} \mid \mathbf{Y}) \leq \varsigma/(1 + K)$ as in equation (8). Note that the specification of a cost ς may not be easy in practice. As an alternative, one may place a cap on the number of network relationships, say n^*, that a biologist may look at in future experiments. Another option is to specify a cost ς that explicitly controls the expected posterior FDR. This requires specifying a null hypothesis H_0 and an alternative H_1 in relation to what may be considered a meaningful relationship. Let $H_{0ik} : D_{ik} \geq \gamma$ and $H_{1ik} : D_{ik} < \gamma$, for each pair $i \neq k$, where γ denotes a timing envelope of interest. Clearly, using the notation of Section 2.2.1, we can define p_{ik} as the posterior probability $P(D_{ik} \geq \gamma \mid \mathbf{Y}) \approx 1/M \sum_{j=1}^{M} I(D_{ik}^{(j)} \geq \gamma)$ and proceed by selecting the optimal cost ς^* as:

$$\varsigma^* = (1 + K)\, E\big(D_{ik}^{\ell} \,\big|\, \mathbf{Y}\big), \tag{13}$$

where $\ell = \sup(q : \sum_{j=1}^{q} p_{ik}^q < q\alpha)$ and p_{ik}^q ordered so that $E(D_{ik}^1 \mid \mathbf{Y}) \leq E(D_{ik}^2 \mid \mathbf{Y}) \leq \ldots \leq E(D_{ik}^C \mid \mathbf{Y})$, where $C = C_2^N$.

The above approach recognizes the importance of the timing characteristics of gene expression. The selection of an appropriate timing envelope γ must, however, be aided by biological knowledge about the timing of gene–gene regulation in the specific process under investigation. For example, in cell cycle experiments, regulatory envelopes of interest may span only a few minutes (Spellman et al., 1998), while in the study of androgen refractory tumors the timing of interest is of the order of days (Pound et al., 1999).

3. Applications

In this section, we apply our model to a set of simulated data and to time course microarray data arising from animal studies on prostate cancer progression. Our inferences are based on 15,000 samples from the posterior distribution of the model parameters obtained after discarding the initial 20,000 MCMC iterations for burn-in.

3.1 *Simulation*

Let $y_i(t) = a_i\, f(t + \delta_i) + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ and $\delta_i \overset{iid}{\sim} U[-1, 1]$. Moreover, assume that the functional mean $f(t)$ takes one of the following five generating forms:

$$\begin{aligned} f_1(t) &= -[\sin\{(t + 0.5)/4\} + \cos\{(t - 1)/5\}], \\ f_2(t) &= \cos(t/4), \\ f_3(t) &= \sin\{(t + 0.5)/4\} + \cos\{(t - 1)/5\}, \\ f_4(t) &= -\cos(t/4), \\ f_5(t) &= \sin(t/6). \end{aligned}$$

Assuming that $\sigma_\epsilon = 0.4$ and that $a_i \sim N(1, 0.2)I(a_i > 0)$, we simulated trajectories for 40 pseudogenes over 30 equally spaced time points in the interval $T = [0, 30]$ from each of the above functions, in order. Additionally, we added 300 "nondifferentially" expressed pseudogenes simulated from $N(c_i, \sigma_\epsilon^2)$, with $c_i \sim U(-1, 1)$.

We note that the 500 pseudogenes are not simulated from our model. In fact, here we use five different shape functions, with different levels of synchronicity and different numbers of functional features (local extrema) over the time domain.

We model the common shape function with 30 equally spaced interior knots and the time-transformation functions with three equally spaced knots (see Section 2.1.2 for considerations about these choices). We also consider a maximum expansion constraint $\Delta = 5$.

Panels (a) and (b) of Figure 2 show, respectively, the simulated and fitted (posterior mean) profiles. Panel (c) shows the expected posterior amplitude values. The first 200 trajectories are successfully classified as belonging to the overly active class. Controlling the expected posterior FDR at the level 0.05 we select 210 pseudogenes with no false negatives (panel (d)). Our selection is similar to that obtained when applying the method of Storey (2007) (See Web Supplementary Materials, Section 3).

Panel (e) shows the median time-transformation functions. We note that the time-transformation functions clearly identify the three patterns of synchronicity used to generate the pseudogenes. Panel (f) shows the expected posterior global distances between each pair of pseudogenes. In the resulting matrix, darker areas represent smaller distances, and thus stronger associations. The chess-like pattern in the association matrix shows that we successfully identified within-curve similarities of trajectories generated from the same functional mean $f_k(t)$ $(k = 1, \ldots, 5)$ and between-curve similarities between pseudogenes simulated from $f_1(\cdot)$,$f_3(\cdot)$ and $f_2(\cdot)$,$f_4(\cdot)$, which reflects the functional relationships $f_1(t) = -f_3(t)$ and $f_2(t) = -f_4(t)$. The lighter shade of gray associated with the last functional class $f_5(t)$ as related to profiles generated from $f_1(t)$ and $f_3(t)$ reflects that these profiles achieve synchronicity only over a partial section of the time domain. The degree of posterior separation between pseudogenes that are not supposed to be related (lightly colored versus dark colored areas in the matrix) is in general very well defined. In the Web Supplementary Materials, Section 4, we compare the results from our model to those obtained using the Gaussian partial correlation method implemented in the `R` package `GeneNet` (Opgen-Rhein and Strimmer, 2006b). Our inferences using the posterior mean distances offer a sharper identification of the patterns of synchronicity when compared to inferences obtained using partial correlation estimates from `GeneNet`.

We also examined sensitivity of the results to the choice of the parameter σ_ϵ. Our analyses (Web Supplementary Materials, Section 5) indicate that our model still gives a good separation between unrelated genes when profiles are simulated with increased variability.

3.2 *Case Study*

3.2.1 *Background.* The diagnosis and treatment of prostate cancer have changed dramatically over the last 20 years parallel to an increased understanding of the natural history of the disease. As a result of these advances, use of androgen withdrawal therapies has grown as an effective way to slow down prostatic neoplasms proliferation. Although the majority of tumors regresses in response to androgen ablation therapy, almost all eventually progress to a state of androgen independence, characterized by tumor growth despite the androgen-depleted environment.

The Shionogi tumor model is an androgen-dependent model of mouse origin. Because patterns of change in gene expression after castration of the animals are similar to those seen in humans, this model has been validated as a model for human disease.

In this analysis, we utilize data from 6- to 8-week-old mice implanted with Shionogi xenografts and castrated at day 14 post implantation. Shionogi tumor cells were isolated at different time points: precastration (day 0) and from day 1 to 25 postcastration with mRNA obtained for microarray analysis. The sampling design consists of 17 mRNA expression measurements per gene, collected at unequally spaced time points between day 0 and day 25. For this application we consider 2357 genes.

Data were preprocessed and normalized using methods implemented in the `R`-package `Limma` from Bioconductor.

3.2.2 *Analysis and results.* Figure 3 shows the data and the results from fitting our model. Specifically, panel (a) shows mRNA time course expression profiles for a random sample of genes. Panel (b) shows the posterior mean of the amplitude parameters, $E(a_i|\mathbf{Y})$, versus the posterior mean probabilities of normal expression, $E(\pi^0|\mathbf{Y})$. Applying the method discussed in Section 2.2.1 to the posterior samples of the amplitude parameters, controlling the posterior expected FDR at the 0.01 level, we selected a set of 456 differentially expressed genes for network analysis. Panels (c)–(f) show a sample of gene-expression profiles superimposed with the posterior mean mRNA abundance profiles and simultaneous 95% credible bands.

Figure 4 shows the results from our network analysis over the set of 456 differentially expressed genes. Panel (a) shows the (transformed) posterior mean global distances (i.e., $E[\exp\{-D_{ik}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_k)\} \mid \mathbf{Y}]$), against the posterior probability of the average timing distance being at least one day (that is, $P\{D_{ik}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_k) \geq 1 \mid \mathbf{Y}\}$). The vertical line in panel (a) shows the decision boundary, controlling the expected posterior FDR for the network relationships at the level 0.05. Similarly, panel (b) shows the expected posterior FDR versus the number of differential network relationships. The horizontal line corresponds to the boundary controlling the expected posterior FDR at 0.05. Panel (c) shows the corresponding gene–gene expected posterior global distance matrix (genes were ordered to visualize possible interaction structures using the `R` package cluster). Finally, panel (d) shows the set of interactions selected to control the expected posterior FDR at level $\alpha = 0.05$. The presence of a significant network relationship between genes i and k is pictured as a dark spot in the (i, k) entry of the matrix in panel (d).

After castration, androgen levels in mice are virtually reduced to zero and tumor cells undergo apoptosis leading to tumor regression. However, after an initial phase of induced apoptosis, lasting approximately 7 days, tumor cells become androgen-independent and they start to grow. Thus, it may prove useful to look at how genes interact with each other during different phases of the biological process under study. We consider the changes in gene–gene regulatory networks up to 7 days and between 7 and 25 days after castration. We build the networks on slightly modified local measures where we take average distances over the two time periods. Figure 5 shows changes in the cluster structure of the distance matrix and associated changes in the topology of the inferred network.

In order to interpret the biological information captured by our network analysis, we looked at a subset of transcription regulators and genes with known pairwise relationships related to regulation of expression in the ingenuity database. Table 1 shows the subset of genes with significant interactions

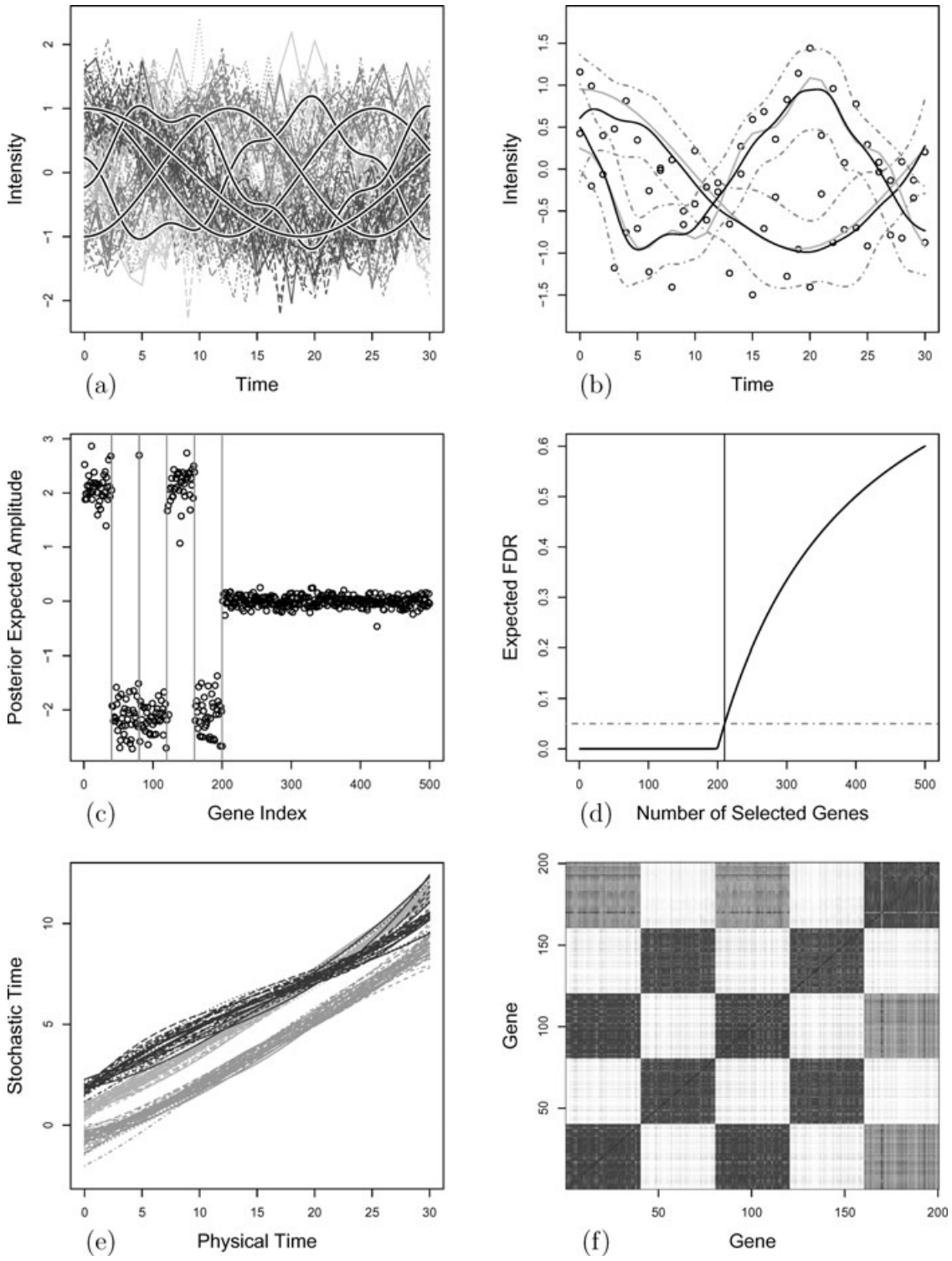

Figure 2. Simulation study. (a) Simulated pseudogene trajectories superimposed with true shape functions (solid lines). (b) Fitted median profiles (solid black) for a random sample of pseudogenes along with 95% credible interval (dot–dashed lines) superimposed with true signal (solid gray). (c) Expected posterior amplitudes $E(a_i \mid \mathbf{Y})$. (d) Expected posterior FDR versus number of selected genes. (e) Posterior median time-transformation functions. (f) Gene–gene expected posterior global distance matrix.

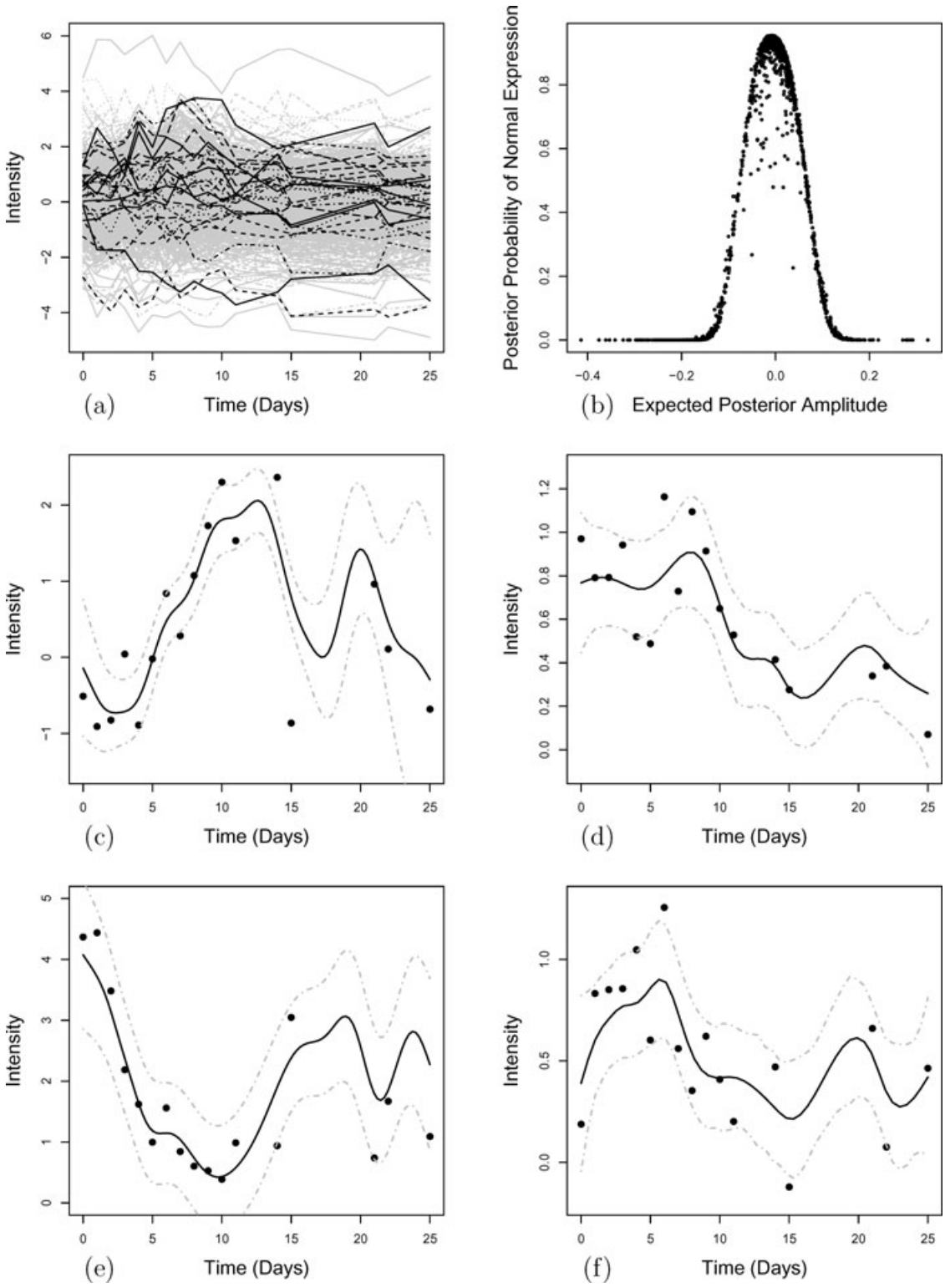

Figure 3. Case Study. (a) Gene-expression profiles. (b) Posterior mean amplitude versus the posterior mean probability of normal expression. (c)–(f) Posterior mean profiles (solid line) for a sample of four genes superimposed with simultaneous 95% credible bands. Dots represent the observed data points.

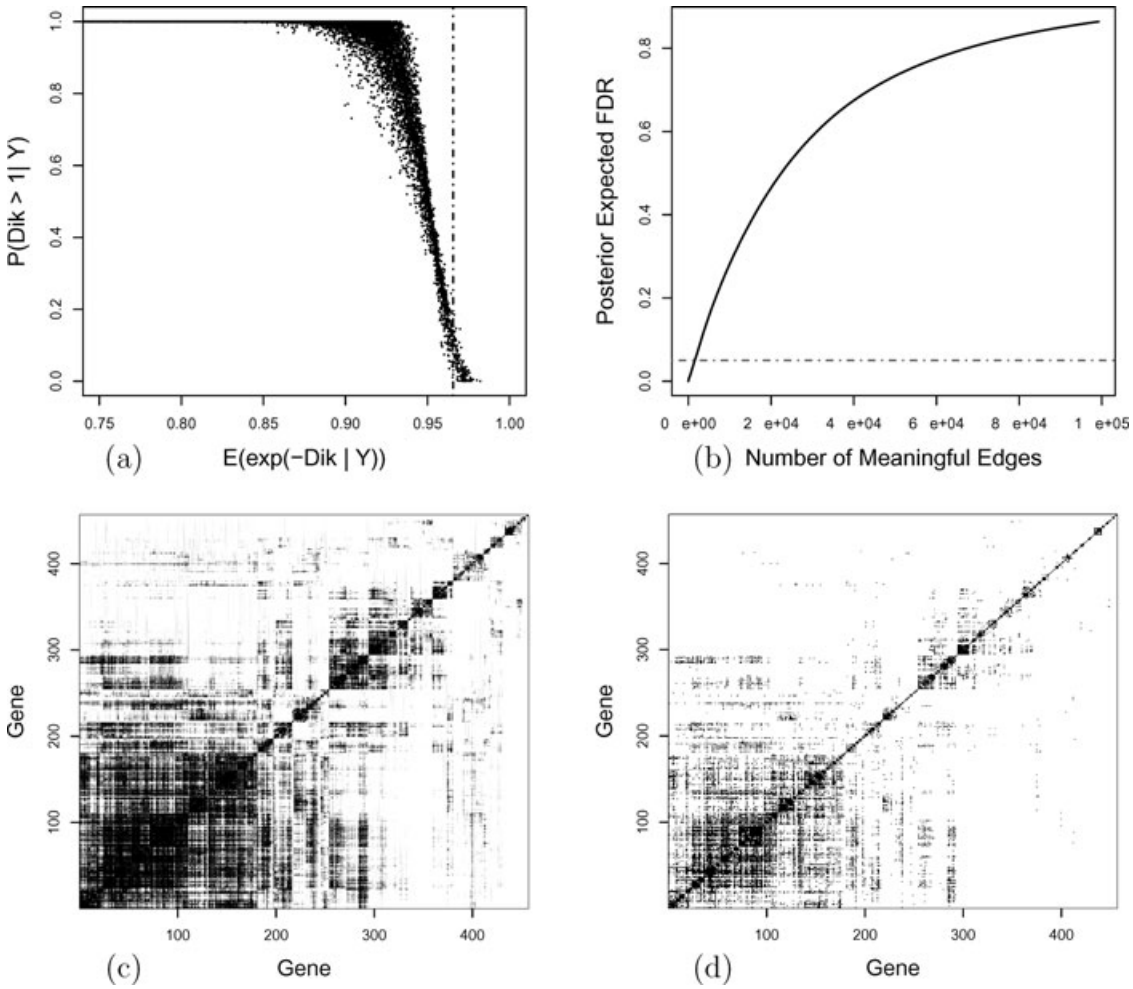

Figure 4. Case study. (a) Expected posterior global distance versus $P(D_{ik}(u_i, u_k) > 1 \mid \mathbf{Y})$ with decision boundary controlling the expected posterior FDR at level 0.05. (b) Expected posterior FDR by number of differential interactions. (c) Expected posterior global distance matrix (darker areas indicate higher synchronicity). (d) Global network associated with the distance matrix in (c) (dark spots correspond to the edges selected in (a)).

(posterior probability less than or equal to 0.05 according to our analysis). In the table, genes under the first column are transcription regulators. Analysis of the selected network with Cytoscape software (`http://www.cytoscape.org/`) revealed the presence of six subnetworks related to biological processes relevant to our system. Specifically, two subnetworks (subnetworks 1 and 5) may be related to T-cell infiltration of tumors that occurs in the Shionogi model upon castration of mice (Nesslinger et al., 2007). Genes in Sub1 (SPP1, SPI1, EMR1, ELA2, CSF1R) are related to proliferation, apoptosis, and differentiation of leukocytes as well as chemotaxis of leukocytes. Moreover, genes in Sub5 (APEX1, HMGB2, SET) are part of the 'Granzyme A mediated Apoptosis Pathway' according to BIOCARTA (`http://www.biocarta.com/`). Thus, it is possible that in our system, infiltrating T-lymphocytes result in the release of Granzyme A in Shionogi tumor cells, leading to an additional activation of caspase-independent apoptosis pathway. Genes in Sub2 (RUNX1T1, CD53, OMD, EZH2, SERPINF1, JUND, HCK) are mainly related to cell proliferation and apoptosis. Genes in Sub3 (PSMA2, NFE2L2, PSMA6, PSMA5, SOD2) are related to the ubiquitin proteasome pathway and oxidative stress. The ubiquitin proteasome pathway has an important role in the degradation of proteins. This oxidative pathway combats the accumulation of reactive oxygen containing molecules that are produced in the cell in response to stress. Levels of oxidative stress affects the effectiveness of radiotherapy and severe oxidative stress can damage DNA and proteins and trigger apoptosis. In Sub4, genes NEUROG3 and PAX6 are related to differentiation of neurons. In the context of prostate cancer progression there is an increase in cells with a neuroendocrine phenotype following androgen ablation and it is thought that the neuropeptide hormone produced from these cells may impact on tumor biology (Amorino and Parsons, 2004) and that NEUROG3 is expressed in metastatic neuroendocrine prostate cancer cells (Hu et al., 2002). Finally, the two genes in Sub6 (MTPN, NPPB) are related to apoptosis and their relationship is supported in the Ingenuity database.

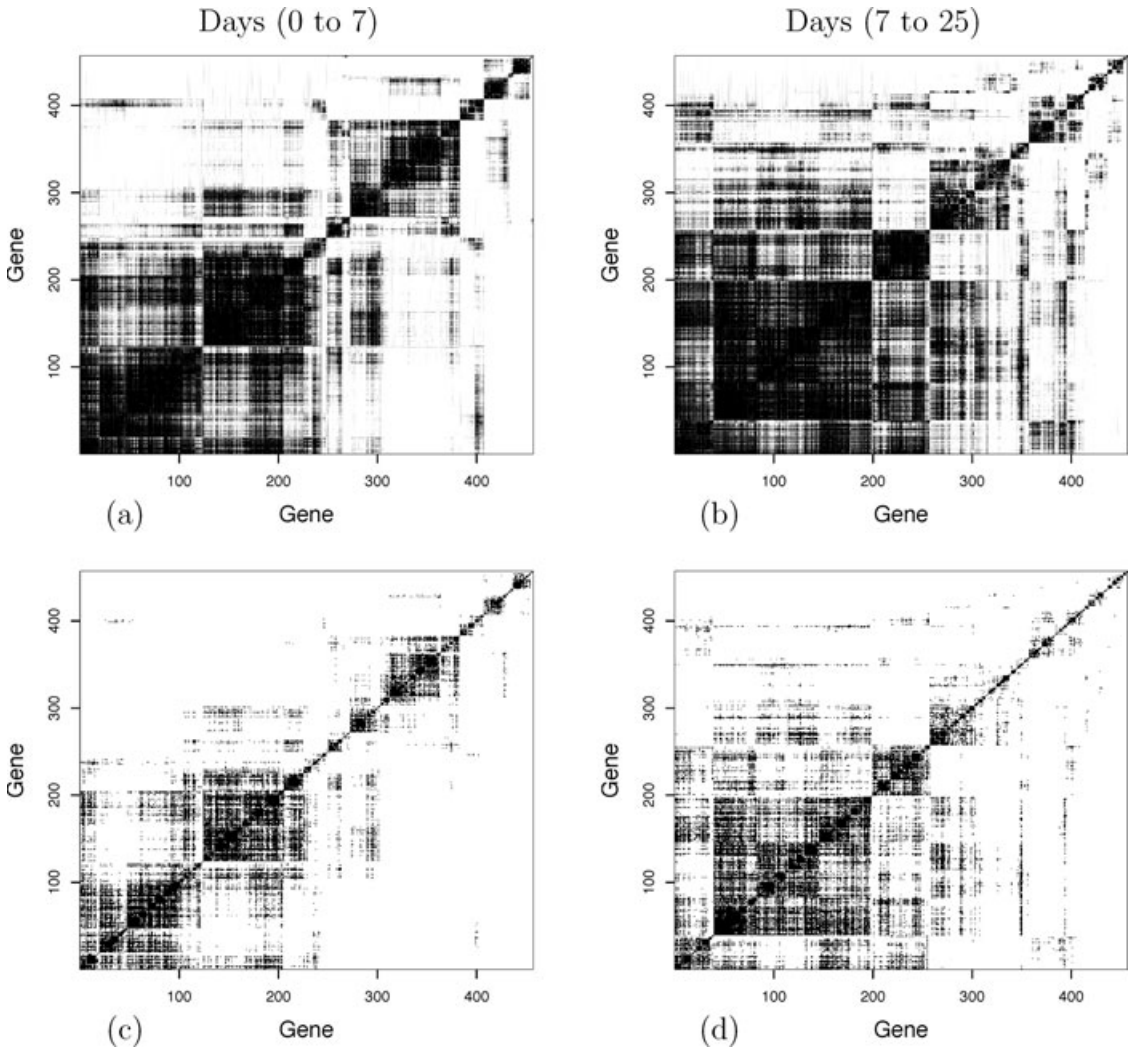

Figure 5. Case study. (a) Local timing distance matrix (days 0 to 7). (b) Local timing distance matrix (days 7 to 25). For both panels, darker areas correspond to higher levels of synchronicity. (c)–(d) Dark spots correspond to relationships selected to control the expected posterior FDR at a level $\alpha = 0.05$.

4. Discussion

In this article, we propose a model-based framework for selecting differentially expressed genes and inferring gene network relationships based on the characterization of profile similarities of time course microarray data. Our model assumes that variation of gene-expression profiles can be sufficiently well captured by gene-specific linear transformations of a common shape function evaluated over a gene-specific stochastic time transformation. We showed that our method is flexible enough to fit even profiles that violate the assumption of a common shape function (Section 3.1). Moreover, we showed that our model validates biologically significant relationships that are plausible based on the current literature (Section 3.2). The approach outlined in this article is likely to work well when considering time series long enough to allow for the identification of a functional response.

Differential expression in the time course setting has been previously defined as a significant variation of the mRNA abundance signal over time (Angelini et al., 2007; Storey, 2007). In this article, we adhere to this concept, proposing a model-based framework for the definition of abnormal activity in gene expression. We base our inferences on the estimated amplitude parameters indicating the strength of the mRNA abundance signal.

Assessing regulatory relationships between genes based on the level of synchronicity of their expression profiles has also been considered by other investigators (see, e.g., Qian et al., 2001; Leng and Müller, 2006). In contrast to these previous approaches, our method does not depend on equally spaced sampling time points. Moreover, our model allows for time shifts but also nonlinear transformations in the gene-specific time scales, making our representation suitable to the analysis of expression profiles exhibiting more than one functional feature over the sampling design interval.

The focus of this article is on utilizing a model-based framework that allows for inferences on both differential expression

Table 1
Biological interpretation of the network in a subset of genes where relationships are related to regulation of expression

Gene 1	Gene 2	$P(D_{ij} \geq 1\,\text{day} \mid Y)$	Notes
Sub 1			
SPI1	SPP1	0.039	Proliferation, apoptosis
SPI1	ELA2	0.001	and differentiation
SPI1	CSF1R	<0.001	of leukocytes
SPI1	EMR	<0.001	
Sub 2			
RUNXIT1	SERPINF1	<0.001	Cell proliferation
RUNXIT1	OMD	0.005	and apoptosis
RUNXIT1	CD53	<0.001	
RUNXIT1	EZH2	0.016	
RUNXIT1	HCK	<0.001	
RUNXIT1	JUND	<0.001	
Sub 3			
NFE2L2	PSMA2	<0.001	Ubiquitin
NFE2L2	PSMA5	<0.001	proteasome pathway
NFE2L2	PSMA6	0.029	
NFE2L2	SOD2	<0.001	
Sub 4			
PAX6	NEUROG3	0.013	Neuronial
PAX6	EHBPIL1	<0.001	differentiation
Sub 5			
HMGB2	SET	0.01	Granzyme apoptosis
HMGB2	APEX1	<0.001	pathway
Sub 6			
MTPN	NPPB	0.004	Apoptosis

and network relationships. To our knowledge, no previous work has addressed these two tasks simultaneously. Even so, we compared our approach with single-tasks approaches. Using a simulation study (Web Supplementary Materials, Section 3) we compared our approach with that proposed by Storey (2007). We showed that our method selects a similar set of genes. We also compared our approach for inferring network relationships with that proposed by Opgen-Rhein and Strimmer (2006b) (Web Supplementary Materials, Section 4) and showed that our method identifies relationships missed by `GeneNet`.

We note that our results are mostly dependent on gene-expression data because our priors are fairly diffuse. Additional prior structure related to the biological knowledge of existing genetic interactions may improve the quality of our inferences and could, in principle, be integrated in our model via a conditional independence prior at the level of the time-transformation coefficients $\boldsymbol{\phi}$ and scale parameters $(\mathbf{c}, \mathbf{a})$. This would, however, increase the model complexity from linear to combinatorial in the number of genes.

5. Supplementary Materials

Web Tables and Figures referenced in Sections 2.1.1, 2.1.2, and 3.1 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We acknowledge support by grants 1P50CA097186-019002 and 1P50CA097186-010003 from the National Cancer Institute. LI also acknowledges partial support from the Career Development Funding from the Department of Biostatistics, University of Washington.

References

Allocco, D., Kohane, I. and Butte, A. (2004). Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics* **5,** 1–10.

Amorino, G. and Parsons, S. (2004). Neuroendocrine cells in prostate cancer. *Critical Review of Eukaryotic Gene Expression* **14,** 287–300.

Angelini, C., De Canditiis, D., Mutarelli, M., and Pensky, M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **6,** 1–33.

Beal, M., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* **21,** 349–356.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57,** 289–300.

Bratsun, D., Volfson, D., Tsimring, L., and Hasty, J. (2005). Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **102,** 14593–14598.

Brumback, L. C. and Lindstrom, M. J. (2004). Self modeling with flexible, random time transformations. *Biometrics* **60,** 461–470.

Chi, Y., Ibrahim, J., Bissahoyo, A., and Threadgill, D. (2007). Bayesian hierarchical modeling for time course microarray experiments. *Biometrics* **63,** 496–504.

de Boor, C. (1978). *A Practical Guide to Splines.* Berlin: Springer-Verlag.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–102.

Gervini, D. and Gasser, T. (2004). Self-modelling warping functions. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **66,** 959–971.

Hu, Y., Ippolito, J., Garabedian, E., Humphrey, P., and Gordon, J. (2002). Molecular characterization of a metastatic neuroendocrine cell cancer arising in the prostates of transgenic mice. *Journal of Biological Chemistry* **277,** 44462–44474.

Inoue, L., Neira, M., Nelson, C., Gleave, M., and Etzioni, R. (2007). Cluster-based network model for time course gene expression data. *Biostatistics* **8,** 507–525.

Leng, X. and Müller, H. (2006). Time ordering of gene co-expression. *Biostatistics* **7,** 569–584.

Markowetz, F. and Spang, R. (2007). Inferring cellular networks—a review. *BMC Bioinformatics* **8** (Suppl 6), S5.

Michalak, P. (2008). Coexpression, coregulation, and cofunctionality of neighbouring genes in eukaryotic genomes. *Genomics* **91,** 243–248.

Morris, J., Brown, P., Baggerly, K., and Coombes, K. (2006). Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. *Bayesian Inference for Gene Expression and Proteomics*, K. A. Do, P. Mueller, and M. Vannucci (eds). New York: Cambridge University Press.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association* **99,** 990–1001.

Müller, P., Parmigiani, G., and Rice, K. (2006). FDR and Bayesian multiple comparisons rules. *Proceedings of the Valencia/ISBA 8th World Meeting on Bayesian Statistics*. Oxford: Oxford University Press.

Nesslinger, N. J., Sahota, R. A., Stone, B., Johnson, K., Chima, N., King, C., Rasmussen, D., Bishop, D., Rennie, P. S., Gleave, M., Blood, P., Pai, H., Ludgate, C., and Nelson, B. H. (2007). Standard treatments induce antigen-specific immune responses in prostate cancer. *Clinical Cancer Research* **13,** 1493–1502.

Newton, M. A., Noueiry, A., Sarkar, D., and Alquist, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5,** 155–176.

Opgen-Rhein, R. and Strimmer, K. (2006a). Inferring gene dependency networks from genomic longitudinal data: A functional data approach. *REVSTAT* **4,** 53–65.

Opgen-Rhein, R. and Strimmer, K. (2006b). Using regularized dynamic correlation to infer gene dependency networks from time-series microarray data. *Proceedings of the 4th International Workshop on Computational Systems Biology, WCSB 2006*, Tampere, Finland.

Parmigiani, G., Garrett, S. E., Anbashgahn, R., and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer. *Journal of The Royal Statistical Society, Series B* **64,** 717–736.

Peña, J. (1997). B-splines and optimal stability. *Mathematics of Computation* **66,** 1555–1560.

Pound, C. R., Partin, A. W., Eisenberger, M. A., Chan, D. W., Pearson, J. D., and Walsh, P. C. (1999). Natural history of progression after PSA elevation following radical prostatectomy. *Journal of the American Medical Association* **281,** 1591–1597.

Qian, J., Dolled-Filhart, M., Lin, J., Yu, H., and Gerstein, M. (2001). Beyond synexpression relationships: Local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *Journal of Molecular Biology* **314,** 1053–1066.

Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **60,** 351–363.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell* **9,** 3273–3297.

Storey, J. D. (2007). Significance analysis of time course microarray experiments. *PNAS* **101,** 12837–12842.

Telesca, D. and Inoue, L. Y. T. (2008). Bayesian hierarchical curve registration. *Journal of the American Statistical Association* **103,** 328–339.

Weber, W., Kramer, B., and Fussenegger, M. (2007). A genetic time-delay circuitry in mammalian cells. *Biotechnology and Bioengeneering* **98,** 894–902.

Received April 2007. *Revised June* 2008.
Accepted June 2008.